**A8Wa Chi-square goodness-of-fit test**

In this section we will explore the concept of measuring how well a data set can be modelled by a probability distribution. The method is called the **chi-square goodness-of-fit test.** This test makes use of the chi-square distribution to compare the observed frequencies (O) with the expected frequencies (E). The expected frequencies are predicted by fitting a probability distribution to the data set of observed frequencies. The objective of comparing the observed sample frequencies with expected frequencies is to establish if they differ significantly.

If you research this test, you will establish that the chi-square test is an alternative to the Anderson-Darling and Kolmogorov-Smirnov goodness-of-fit tests. The chi-square goodness-of-fit test can be applied to discrete distributions such as the binomial and the Poisson. The Kolmogorov-Smirnov and Anderson-Darling tests are restricted to continuous distributions.

For a chi-square goodness of fit test, the hypotheses take the following form:

H0: The data are consistent with a specified distribution

H1: The data are not consistent with a specified distribution

The goodness-of-fit can then be assessed by conducting a chi-square test on the observed and expected frequencies as defined by equation (1).

$χ\_{cal}^{2}= \sum\_{i=1}^{n}\frac{\left(O\_{i}- E\_{i}\right)^{2}}{O\_{i}}$ (1)

Where the degrees of freedom (df) are given by equation (1)

df = n - k - 1 (2)

Where n = number of categories or classes remaining after combining classes, and k = number of population parameters of a distribution that must be estimated to calculate the expected frequencies. Examples of values of k are for different probability distributions are described as follows:

(a) Comparing an observed frequency against a known expected frequency, k = 0

df = n – 1

(b) For a uniform distribution, k = 0

df = n – 1

(c) For a normal distribution, X~N (, 2)

|  |  |
| --- | --- |
| Value of k | Parameter conditions |
| 0 |  and  are known |
| 1 |  known and  estimated |
| 1 |  estimated and  known |
| 2 |  and  are estimated |

Table 1

(d) For a binomial distribution, X~B (n, p)

|  |  |
| --- | --- |
| Value of k | Parameter conditions |
| 0 | p known |
| 1 | p estimated |

Table 2

(e) For a Poisson distribution, X ~P ()

|  |  |
| --- | --- |
| Value of k | Parameter conditions |
| 0 |  known |
| 1 |  estimated |

Table 3

****

**Example 1**

To illustrate the method, consider the example of a motorway safety officer. The officer believes that the number of purchases per week that a typical family in Region A makes for a product Q can be modelled using a Poisson distribution. If X denotes the number of purchases per week then the sample data can be modelled by fitting a Poisson distribution to the sample data.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Number of purchases, X | 0 | 1 | 2 | 3 | 4 | 5 | 6 |
| Observed frequency, O | 16 | 33 | 45 | 32 | 26 | 18 | 11 |

Table 7.21

**The Poisson probability distribution is given by equation (4.29).**

$$P\left(X=r\right)= \frac{λ^{r} e^{- λ}}{r!}$$

Where r = 0, 1, 2, 3, ............, .

The five-step procedure to conduct this test progresses as follows.

**Step 1 - State hypothesis**

H0: The data are consistent with a Poisson distribution

H1: The data are not consistent with a Poisson distribution

**Step 2 - Select test**

Comparing observed frequency with an expected frequency predicted by the Poisson distribution. Chi square goodness-of-fit test.

**Step 3 - Set the level of significance** () (see Cell L12)

**Step 4 - Extract relevant statistic**

The initial calculation process involves using the sample data to estimate the average number of accidents per week, . Given that we have a frequency distribution then we can calculate the average value as follows.

|  |  |  |
| --- | --- | --- |
| Number of purchases, X | Observed frequency, f | X \* f |
| 0 | 16 | 0 |
| 1 | 33 | 33 |
| 2 | 45 | 90 |
| 3 | 32 | 96 |
| 4 | 26 | 104 |
| 5 | 18 | 90 |
| 6 | 11 | 66 |
| Totals = | 181 | 479 |

Table 2

**The mean for a frequency equation is given by equation ()**

$$λ= \frac{\sum\_{i=1}^{n}X\_{i}f\_{i}}{\sum\_{i=1}^{n}f\_{i}}$$

From table 7.22, ∑ Xf = 0 + 33 + …. + 66 = 479, ∑ f = 16 + 33 + ….. + 11 = 181

$$Estimated mean λ= \frac{479}{181}=2.6464$$

The individual Poisson probabilities for X = 0, 1, 2,…, and 6 can now be calculated using equation (4.29). For example:

$$P\left(X=0\right)= \frac{2.6464^{0} e^{- 2/6464}}{0!}=0.0709$$

$$P\left(X=1\right)= \frac{2.6464^{1} e^{- 2.6464}}{1!}=0.1876$$

Table 7.23 illustrates the probability calculation values for X = 2, 3, 4, 5, and 6. If a Poisson distribution fits the observed sample frequency data (O) then we can use equation (3)

$E\left(X=r\right)= \left(\sum\_{i=1}^{n}f\_{i}\right) ×P(X=r)$ (3)

to calculate what the expected frequencies (E) would be if the Poisson distribution is appropriate. Finally, we wish to find out whether the observed and expected frequencies are close enough for us to accept the null hypothesis. This is achieved by undertaking a chi-square test to compare these two frequency values using equation (3).

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Number of purchases, X | Observed frequency, f | X \* f | P(X) | E | (O – E)2/E |
| 0 | 16 | 0 | 0.0709 | 12.8339 | 0.7811 |
| 1 | 33 | 33 | 0.1876 | 33.9637 | 0.0273 |
| 2 | 45 | 90 | 0.2483 | 44.9409 | 0.0001 |
| 3 | 32 | 96 | 0.2190 | 39.6440 | 1.4739 |
| 4 | 26 | 104 | 0.1449 | 26.2286 | 0.0020 |
| 5 | 18 | 90 | 0.0767 | 13.8823 | 1.2214 |
| 6 | 11 | 66 | 0.0525 | 9.5067 | 0.2346 |
| Totals = | 181 | 479 |  |  | 3.7403 |

Table 3

The value of P(X = 6) = 1 – P(X ≤ 5) = 0.0525.

The value of the chi-square test statistic =3.7403.

For the Poisson distribution we estimated  using the sample data. Given this estimation we have a value for k of 1 and the number of degrees of freedom df = n – k – 1 = 7 – 1 – 1 = 5.

**Step 5 - Make decision**

The calculated chi-square test statistic = 3.7403. The critical chi-square value depends upon both the level of significance ( = 0.05) and the number of degrees of freedom (df = 5). The critical value can be found from statistical tables to give the critical chi-square critical value equal to 11.0705. Does the test statistic lie within the region of rejection? Compare the calculated and critical values to determine which hypothesis statement (H0 or H1) to accept. From Excel, calculated chi-square value < critical chi-square value (3.7403 < 11.0705), and we would fail to reject the null hypothesis H0 and reject the alternative hypothesis H1.

Figure 1 illustrates the relationship between the test statistic, critical test statistic, and p-value.



Figure 1

Conclude that the there is a significant relationship between the observed and expected frequencies. This implies that the data can be modelled by a Poisson probability distribution with an estimated average value of 2.6464.

Note that the expected number for each random variable must be at least 5. If necessary, combine classes in the table to satisfy this requirement or redo the data collection and increase the size of the sample.

****

**Excel solution**

Figures 2 and 3 illustrate the Excel solution.



Figure 2

**Excel solution**

X Cells B5:B11 Values

O Cells C5:C11 Values

xO Cells D5 Formula:=B5\*C5

 Copy formula down D5:D11

∑f = Cell C12 Formula:=SUM(C5:C11)

∑fX = Cell D122 Formula:SUM(D5:D11)

Estimated mean Cell D14 Formula:=D12/C12

P(X) Cells F5 Formula:=POISSON.DIST(B5,$D$14,FALSE)

 Copy formula down F5:F10

P(X=6) Cell F11 Formula:=1-SUM(F5:F10)

E Cells G5 Formula:=SUM($C$5:$C$11)\*F5

 Copy formula down G5:G11

(O - E)^2/E Cells H5 Formula:=(C5-G5)^2/G5

 Copy formula down H5:H11



Figure 3

**Excel solution**

Significance level Cell L12 Value

Chi square χ2 Cell L15 Formula:=SUM(H5:H11)

n Cell L18 Formula:=COUNT(B5:B11)

k Cell L19 Value = 1

df Cell L20 Formula:=L18-L19-1

P-value Cell L22 Formula:=CHISQ.DIST.RT(L15,L20)

P-value using CHISQ.DIST Cell L23 Formula:=1-CHISQ.DIST(L15,L20,TRUE)

Critical value chi-square Cell L24 Formula:=CHISQ.INV.RT(L12,L20)

Decision:

P-value method

From Excel, the p value = 0.5874 (Cell L24). Does the test statistic lie within the region of rejection? Compare the calculated p-value and test significance level to determine which hypothesis statement (H0 or H1) to accept. From Excel, the p value = 0.5874 (Cell L244) > α = 0.05, and conclude that the evidence suggests that we fail to reject null hypothesis H0 and reject the alternative hypothesis H1.

You could use the Excel function CHISQ.TEST to calculate the p-value but this will give you the p-value when no parameters are estimated (k = 0). This gives a p-value = 0.7118 which as we will see soon agrees with the value given by the SPSS solution.

Critical test statistic method

The calculated chi-square test statistic = 3.7403 (see Cell L15). The critical chi-square value depends upon both the level of significance ( = 0.05, Cell L12) and the number of degrees of freedom (df = 5, cell L20). The critical value can be found from Excel by using the CHISQ.INV.RT() function. The critical chi-square value equals 11.0705 (see Cell L24). Does the test statistic lie within the region of rejection? Compare the calculated and critical values to determine which hypothesis statement (H0 or H1) to accept. From Excel, calculated chi-square value < critical chi-square value (3.7403 < 11.0705), and conclude that the evidence suggests that we fail to reject null hypothesis H0 and reject the alternative hypothesis H1.



**SPSS solution**

SPSS data file Chapter 7\_chi-squared goodness of fit test.sav

Table 4 illustrates the frequency distribution.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| X | 0 | 1 | 2 | 3 | 4 | 5 | 6 |
| OF | 16 | 33 | 45 | 32 | 26 | 18 | 11 |

Table 4

SPSS datafile

Figure 4 illustrates the SPSS data



Figure 4

If we simply type in the number of occurrences for each category, SPSS will think we have seven participants, scoring 16, 33, 45,…….., 11 respectively. We need to force SPSS to treat these values as frequencies for our four categories.

Click on Data > Weight Cases and choose Weight Cases by ObsFreq



Figure 5

Click OK

SPSS will now treat the numbers in the frequency column as the totals for the categories identified in the x column.

Analyse > Non-parametric Test > Legacy Dialogs > Chi-square

Transfer X variable into the Test Variable List box

And in the Expected values box click on Values and copy the expected frequency values into the Values box: 12.8339, 33.9637, 44.9409, 39.6440, 26.2286, 13.8823, and 9.5067. Alternatively, replace the expected frequency values for each x value with the associated probability.



Figure 6

Click on Options

Choose Descriptives



Figure 7

Click Continue

Click OK

SPSS Output

First, you get a table that contains descriptive statistics



Figure 8

Second, you get a table that contains the observed and expected frequencies for each of the powders



Figure 9

Finally, the table shows the Chi-square, the degrees of freedom, and the probability of obtaining this chi-square value merely by chance



Figure 10

SPSS assumes that no parameter values are estimated and therefore calculates the number of degrees of freedom to be equal to 6. For this example, chi-square test statistic = 3.740 and p-value = 0.712. If we changed the Excel solution so that no parameters were estimated, then the number of degrees of freedom = 6 and the p-value = 0.7118. For this example, the chi-square p-value is greater than the significance level (0.712 > 0.05). Thus, we would conclude that our observed frequencies are not significantly different from what we would expect to get by chance. That is, the observed frequency data follows a Poisson probability distribution.

### Check your understanding

X1 A new employment agency has recently implemented a new training programme to develop the interview skills of potential job applicants. Based upon the collected data can we confidently say that the data can be modelled using a binomial distribution (assess at 5%)?

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Number of interview successes | 0 | 1 | 2 | 3 |
| Frequency | 78 | 143 | 43 | 13 |

Table 5

X2 A university has recently set up a satellite department within a local college of higher education. The university claims that 35% of the undergraduate students are in department A, 26% are in department B, 25% are in department D, and 14% are in department D. A random sample of 320 students finds the following number of students in departments A to D: 132, 89, 64, and 35. Perform a hypothesis test at 5% to test this claim.

X3 A new airport terminal has been assessing waiting times for passengers to be processed at the airport check-in counters. The airport owners would like to be able to attach levels of risk to different aspects of the business. To undertake this, we are required to fit an appropriate probability distribution to the observed frequencies provided below.

(a) Use the data in Table 6 to provide an estimate of the population mean and standard deviation,

(b) Construct a z distribution table with upper class boundaries of 14, 17, 22, 26 and infinity,

(c) Use this table to calculate the cumulative distribution function values at these class boundaries based on your answers to part (a) – (c),

(d) Estimate the class probabilities and resultant expected frequencies,

(e) Calculate the observed frequencies based upon your upper-class boundaries, and

(f) Undertake a chi-square goodness-of-fit test to assess at a 95% confidence that the normal distribution would be a good fit to the sample data.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 6 | 7 | 7 | 8 | 10 | 12 | 13 | 13 | 14 |
| 14 | 15 | 15 | 16 | 16 | 16 | 16 | 16 | 17 |
| 17 | 18 | 13 | 18 | 19 | 19 | 19 | 20 | 20 |
| 22 | 23 | 23 | 12 | 24 | 25 | 25 | 26 | 27 |
| 27 | 27 | 28 | 28 | 29 | 30 | 30 | 31 | 33 |

Table 6